



电力业务系统海量 SQL 访问日志预处理方法

李伟伟,张涛,马媛媛,周诚

(国网智能电网研究院,江苏 南京 210003)

摘要:随着智能电网的互动化发展,信息内外网边界数据库交互行为日益增多。为更好地应对电力业务系统数据安全审计的需求,解决数据库审计过程中 SQL 语句容量过大、不易存储和分析的问题,详细设计了一种基于 SQL 模式分析的电力业务系统 SQL 访问日志压缩方法。该方法在不丢失原 SQL 访问语句语义的前提下,实现 SQL 访问日志的压缩,也具有 SQL 日志的简单分类的功能,为数据库的安全审计提供基础。系统测试结果表明,通过该方法可以实现 SQL 访问日志的大幅度压缩,满足电力业务系统数据库行为审计对日志预处理的需求。

关键词:电力;安全审计;日志压缩;预处理

中图分类号:TP311

文献标识码:A

doi: 10.11959/j.issn.1000-0801.2015395

A method of SQL access logs pretreatment for power business system

LI Weiwei, ZHANG Tao, MA Yuanyuan, ZHOU Cheng

State Grid Smart Grid Research Institute, Nanjing 210003, China

Abstract: With the development of smart grid interactive, the database interaction between the information inside and outside is increasing. A method of SQL access log compression for power business system to respond to the demand of database security audit for the power system was designed in detail. This method can realize the compression of SQL access log and have the function of simple classification for SQL logs. The test results show that the method can achieve a large margin compression of SQL access logs and meet the requirements of log preprocessing for the database behavior audit of power business system.

Key words: power business, security audit, logs compression, pretreatment

1 引言

随着智能电网信息化技术和 SG-ERP 的互动化应用开展,电力业务系统不断丰富,内外网交互行为日益多样化。很多业务系统,如电力交易系统、生产系统、财务管控、ERP 系统、营销系统等,都存在着跨内外网边界的数据交

换行为。为了能够及时监控和有效地监督信息外网用户对信息内网各种数据库的访问行为,并及时提醒审计人员对各种违规行为进行审计、追踪和预警,内外网边界数据库交互行为的安全审计^[1]必不可少。内外网边界数据库交互行为审计过程中涉及的信息采集、预处理、数据挖掘和分析等相关技术研究,对智能电网信息安全建设方面有很大

收稿日期:2015-10-21;修回日期:2015-11-31

基金项目:国家电网公司 2015 年科技项目(No. SGRIXTKJ[2015]-614)

Foundation Item: The Science and Technology Project of State Grid Corporation of China in 2015 (No. SGRIXTKJ[2015]-614)

的推动作用。

2 面临的问题和研究现状

目前,公司信息内外网边界的信息交互中,穿透信息内外网边界的数据库访问行为是重要的分析和审计对象,数据库保存着电力业务系统的重要信息,如用户用电信息、营销业务数据等机密、敏感数据。因此,数据库对于黑客和不法人员有着很强的吸引力。数据库建立后会被频繁地使用,频繁地进行插入、修改、删除等操作行为,并且数据库的使用不受时间和地点的限制。因此,数据库的安全性异常重要。由于电力业务系统数据库 SQL 语句操作日志的数据量十分庞大,存储分析存在很大难度。为了更好地对数据库行为继续审计,对电力业务系统 SQL 访问语句进行压缩十分必要。

目前对数据库审计系统^[2-4]的研究已经比较广泛,针对日志压缩相关的研究比较少,2010 年张春玲等人^[5]设计并实现了一种基于虚拟日志压缩的广域网异构数据库异步数据同步方案,虚拟日志压缩通过计算合并日志记录,只保留与数据同步相关的日志信息,从而减少网络数据流量,提高数据同步效率。2010 年王艳峰等人^[6]针对 CN 顶级域名的 DNS 日志从分布式站点传输到数据处理中心时的海量数据存储问题,提出一种高效的 DNS 日志压缩算法,利用 DNS 查询类型的冗余性和 DNS 查询时间、IP 地址和域名等的重复性进行 DNS 日志压缩。

3 SQL 日志压缩方法

同一电力业务系统的数据库操作通过 Web 页面填表方式进行,具有一定的模式可循。SQL 日志压缩方法根据这一特性通过模糊化用户的数字字母等自定义输入,为业务系统创建 SQL 模式库,收到的 SQL 日志通过匹配模式库后仅记录其模式序号即可,大大压缩了 SQL 日志的记录长度,实现 SQL 日志压缩的目的,为下一步的数据库操作行为分析提供便利。

SQL 日志压缩方法如图 1 所示,日志接收服务接收到 SQL 日志,经过扫描程序进行 SQL 语句的解析,形成 SQL 数组,并将无关项替换后,进行 SQL 模式匹配。如果匹配成功,则返回 SQL 模式库的匹配模式,将序号写入 SQL 日志库;如果 SQL 模式库匹配失败,则将 SQL 记录到模式库,同时将序号写入 SQL 日志库。

数据库结构定义如下。

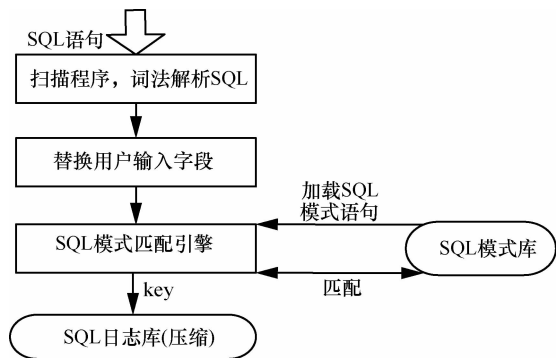


图 1 基于模式匹配的 SQL 日志压缩方法

- 定义应用数据库结构为:{App_ID: 应用序号, App_IP: 应用 IP 地址}。
- 定义 SQL 模式库数据库结构为:{Mode_ID: 模式序号, App_ID: 应用序号, SQL: 模式语句}。
- 定义 SQL 日志库数据库结构为:{SQL_ID: SQL 语句序号, Mode_ID: 模式序号, Time: 时间戳}。

3.1 SQL 词法解析

SQL 词法解析是对存储一段连续缓冲区的 SQL 语句进行分解,得到一个个独立的单词,并且组织成单词链表。对 SQL 进行解析的详细过程如下。

- 遍历 SQL 语句包含的每一个字符,以空白符作为每个单词的开始和结束,解析出 SQL 语句包含的各个单词,同时识别各个单词的类型。目前,规定的单词类型如下:关键字、标识符、数字、数字通配符、字符串通配符、运算操作符、逻辑操作符、分隔符、函数、操作对象、其他。
- 将这些单词存储在单词结构体中,多个单词结构体链接成一个链表。同时根据解析的结果,给单词结构体结构的相关域(如类型)赋值。

SQL 解析过程示例如图 2 所示。

3.2 替换用户输入

由于从应用服务器送出的 SQL 语句是包含用户输入的,而用户输入是不断变化的,如何将正确的 SQL 语句进行抽象,分离出用户输入和 SQL 主干语句是重点。替换规则也是尽可能将用户输入部分提取出来,用某个特定的通配符进行替代,这样就可以将多条结构相似的 SQL 语句抽象为一条 SQL 语句。

替换用户输入示例如图 3 所示。

3.3 SQL 模式匹配引擎

SQL 模式匹配是根据词法解析结果(词法链),匹配

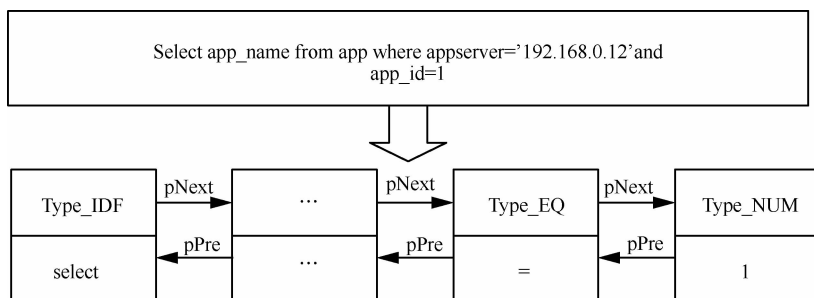


图2 SQL解析过程示例

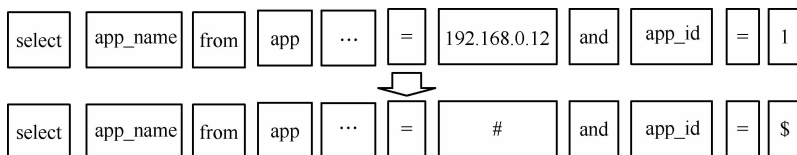


图3 SQL解析后替换用户输入过程示例

SQL 词法链池。详细步骤如下。

- SQL 词法链池对外是一个散列数组，因此匹配过程首先对词法链求散列值，然后在散列数组中进行查找。

散列值计算方法： $\text{Hash} = \text{Hash} \ll 8 + \text{Type}$ （Hash 初始化为 0，Type 为语句词法解析后的单词类型）。

- 通过散列值找到数组对应项，然后在该数组单元下挂词法链表进行比较。检查是否有完全相同的词法链，如果有，则认为 SQL 模式已经存在，仅保存 SQL 模式序号到 SQL 日志表中，否则认为 SQL 模式不存在，保存模式后，再将模式序号保存到 SQL 日志表中。

SQL 模式引擎详细匹配过程如图 4 所示。

4 方法实现

实现电力业务系统 SQL 访问日志的压缩，主要有 3 个模块：初始化模块、模式库匹配模块、结果写入模块。

- 初始化模块：主要是初始化系统的全局互斥变量，通过读取模式库中 SQL 语句，初始化每个业务系统对应的词法扫描链表。
- 模式库匹配：将读取的 SQL 日志通过 SQL 词法解析、用户输入替换，形成链表，与模式库中的 SQL 模式进行匹配。
- 结果写入：对已经存在相应 SQL 模式的日志，只需要写入 SQL 日志表即可，对于没有匹配的日志，则既要写入模式库，又要写入 SQL 日志表。

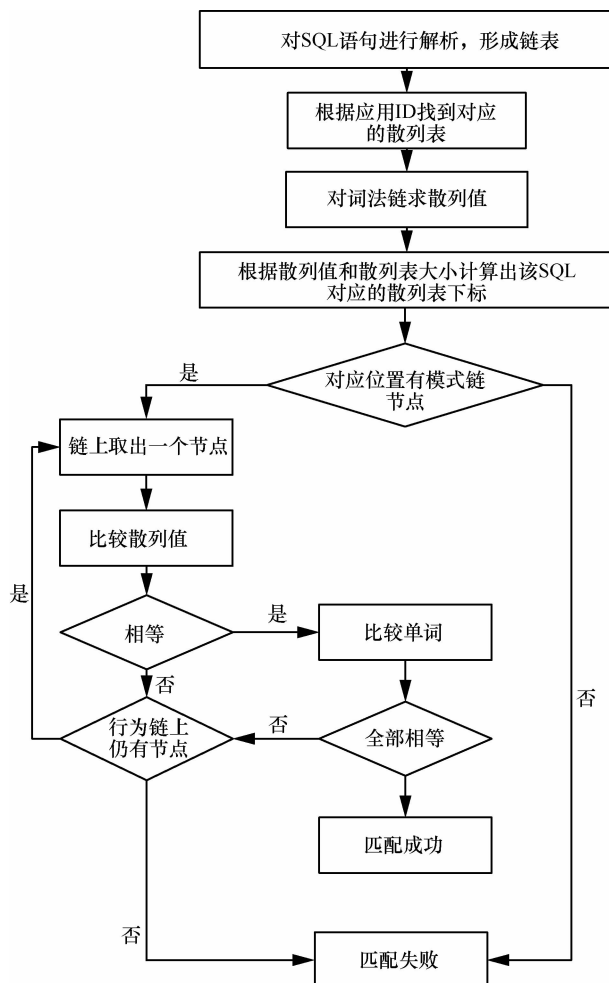


图4 SQL模式库匹配过程

详细的系统实现流程如图 5 所示。

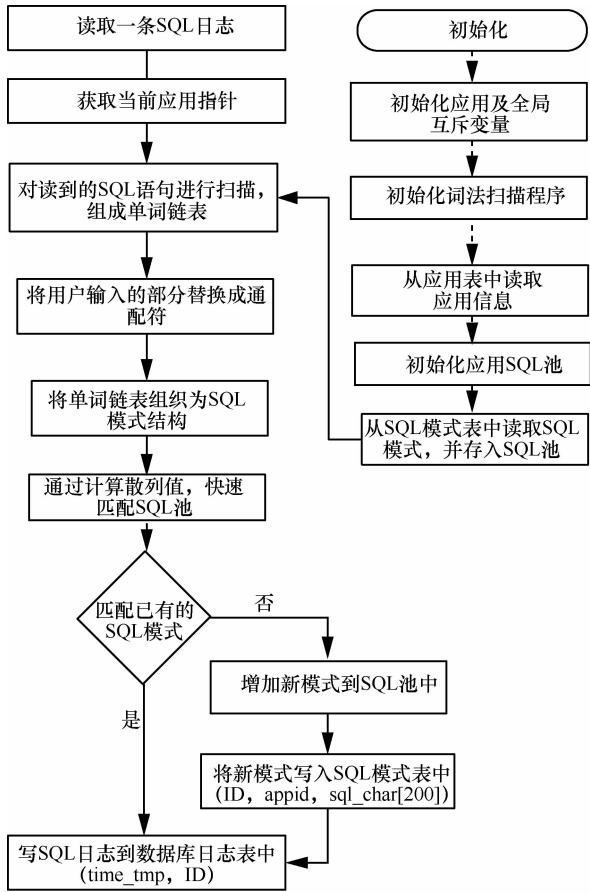


图5 电力业务系统 SQL 访问日志压缩方法实现流程

5 测试结果

5.1 测试环境

实验设备硬件环境: Intel 5500+ICH10R 芯片组, 2 路 4 核 CPU。操作系统是 Windows7。
编程语言: C 语言。

测试库 SQL 语句取样: 业务系统现场采集 SQL 日志。

5.2 测试步骤

- 加载 SQL 日志文本文件;
- 执行压缩, 记录 SQL 模式库中 SQL 语句的条数和 SQL 日志库记录的日志条数。

5.3 测试结果

图 6、表 1 为 SQL 压缩效果测试结果。

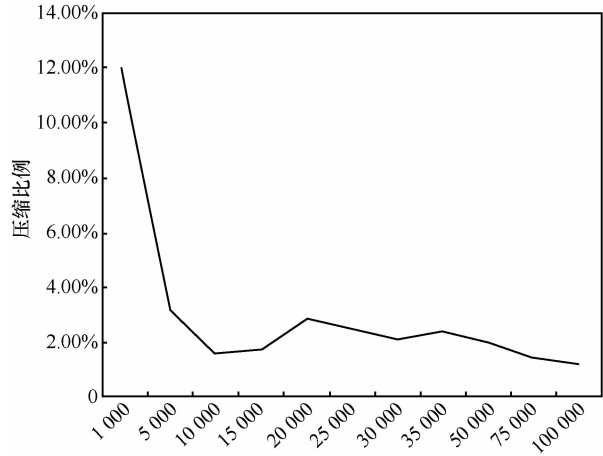


图6 SQL 压缩效果测试结果

5.4 测试结论

通过测试结果可知, 随着业务系统执行的 SQL 数目的增加, 压缩比例达到 50:1。测试过程中, 服务器的 CPU、内存和硬盘各项指标检测正常。

6 结束语

本文设计并实现了一种电力业务系统 SQL 访问日志压缩方法, 该方法除了实现 SQL 访问日志的压缩之外, 也

表 1 SQL 压缩效果测试结果

SQL 日志数/条	SQL 模式库条数/条	SQL 日志库条数/条	压缩比例	失败个数
1 000	119	1 000	11.90%	0
5 000	157	5 000	3.14%	0
10 000	164	10 000	1.64%	0
15 000	261	15 000	1.74%	0
20 000	570	20 000	2.85%	0
25 000	634	25 000	2.54%	0
30 000	647	30 000	2.16%	0
35 000	850	35 000	2.43%	0
50 000	1 024	50 000	2.04%	0
75 000	1 103	75 000	1.471%	0
100 000	1 197	100 000	1.093%	0

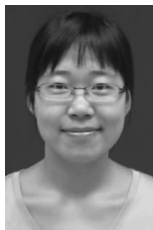


具有 SQL 日志的简单分类的功能,可解决数据库审计过程中 SQL 语句容量过大、不易存储和分析的问题,为电力业务系统安全审计提供了支撑。系统测试结果表明,通过该方法可以实现 SQL 访问日志的大幅度压缩,满足电力业务系统数据库行为审计对日志预处理的需求。

参考文献:

- [1] 方杰,朱京红. 日志挖掘中的数据预处理 [J]. 计算机技术与发展, 2010, 20(4): 17-20.
FANG J, ZHU J H. Data pretreatment of log mining [J]. Computer Technology and Development, 2010, 20(4): 17-20.
- [2] HELMAN P, LIEPINS G. Statistical foundations of audit trail analysis for the detection of computer misuse [J]. IEEE Transactions on Software Engineering, 1993, 19(9): 886-901.
- [3] BISKUP J. Recent advances in intrusion detection [M]. Berlin: Springer, 2000: 28-48.
- [4] 王渊,马骏. 一种基于入侵检测的数据库安全审计 [J]. 计算机仿真, 2007, 24(2): 33-36.
WANG Y, MA J. A method of database secure audit based on intrusion detection[J]. Computer Simulation, 2007, 24(2): 33-36.
- [5] 张春玲,吕震宇,刘遵峰. 基于虚拟日志压缩的数据同步方案[J]. 计算机工程, 2010, 18(36): 67-69.
ZHANG C L, LV Z Y, LIU Z F. Data synchronization solution based on virtual log compression [J]. Computer Engineering, 2010, 18(36): 67-69.
- [6] 王艳峰,王正,阎保平. 一种高效的 DNS 日志压缩算法[J]. 计算机工程, 2010, 15(36): 32-35.
WANG Y F, WANG Z, YAN B P. High-efficient DNS log compression algorithm[J]. Computer Engineering, 2010, 15(36): 32-35.

[作者简介]



李伟伟(1985-),女,国网智能电网研究院工程师,主要研究方向为信息安全。



张涛(1976-),男,国网智能电网研究院高级工程师,主要研究方向为信息安全。



马媛媛(1978-),女,国网智能电网研究院高级工程师,主要研究方向为信息安全。



周诚(1981-),男,国网智能电网研究院工程师,主要研究方向为信息安全。